

EXHIBIT D

Message

From: Melanie Kambadur [REDACTED]@meta.com]
Sent: 8/18/2023 9:52:21 PM
To: Melanie Kambadur [REDACTED]@meta.com]; Nisha Deo [REDACTED]@meta.com]
Subject: Message summary [{"otherUserFbId":100076652037303,"threadFbId":null}]
Attachments: 279159887_704475240903299_7495330466370328987_n.gif

Nisha Deo (8/18/2023 13:34:32 PDT):
 >Hi Melanie - do you know if we used the Books data set in Llama 2?

Melanie Kambadur (8/18/2023 13:34:51 PDT):
 >yes we did

Melanie Kambadur (8/18/2023 13:35:07 PDT):

Redacted

Nisha Deo (8/18/2023 13:36:40 PDT):

Redacted

Melanie Kambadur (8/18/2023 13:38:22 PDT):
 >what do you mean by that?

Nisha Deo (8/18/2023 13:44:01 PDT):
 >WELP for full context we have a inquiry from the atlantic about the use of books3 in Llama -- so basically wondering if it's at all possible for them to deduce that it was used in llama 2

Nisha Deo (8/18/2023 13:44:11 PDT):
 >questions they posed include:

>
 >The research paper about the construction of LLaMa clearly states that Books3 is among the corpora of the program's training data. Can you confirm this is true?
 >
 >- Was Meta aware at the time that Books3 includes pirated books? Do you have further comment about this?
 >
 >- Was Meta aware at the time that Books3 includes books that have been stripped of copyright information, possibly in violation of the DMCA? Do you have further comment about this?
 >
 >- Knowing the above, does Meta have any plans to update LLaMa to remove the use of Books3 or other copyrighted material?

Melanie Kambadur (8/18/2023 13:45:56 PDT):

Redacted

Nisha Deo (8/18/2023 13:47:00 PDT):
 >totally, have looped Ahuva in

Nisha Deo (8/18/2023 13:47:07 PDT):
 >and will start an escalation

Nisha Deo (8/18/2023 13:47:14 PDT):
 >was just trying to understand "ground truth" as they say

Nisha Deo (8/18/2023 13:47:31 PDT):

Redacted

Melanie Kambadur (8/18/2023 13:51:10 PDT):
 >yes I went to an AI event yesterday with a bnch of people from other companies and any time I mention I worked on llama 2 they kept asking me about datasets 😊 I was just like see the paper/ no comment lol

Melanie Kambadur (8/18/2023 13:51:36 PDT):

Redacted

Melanie Kambadur (8/18/2023 13:51:49 PDT):

Redacted

Nisha Deo (8/18/2023 13:55:04 PDT):
 >bafflingly leaky

Nisha Deo (8/18/2023 13:55:16 PDT):

shared: 279159887_704475240903299_7495330466370328987_n.gif

Nisha Deo (8/18/2023 14:04:16 PDT):

>back to books3 -- do you know if assurances from them that their catalog didn't contain any pirated works?

Melanie Kambadur (8/18/2023 14:05:13 PDT):

>I suppose the best defense I can think of is this: <https://github.com/EleutherAI/the-pile/blob/master/LICENSE>

Melanie Kambadur (8/18/2023 14:05:21 PDT):

>the place we pulled it from has a permissive license

Melanie Kambadur (8/18/2023 14:05:29 PDT):

Redacted

Nisha Deo (8/18/2023 14:07:21 PDT):

>totally

Nisha Deo (8/18/2023 14:07:38 PDT):

>Did we have that same license on Llama 1?

Nisha Deo (8/18/2023 14:07:48 PDT):

>(sorry, i owe you multiple drinks for asking you so many questions)

Melanie Kambadur (8/18/2023 14:08:11 PDT):

>no worries. you mean for distributing llama 1? if so, no because llama 1 was a research only model and llama 2 is commercial

Melanie Kambadur (8/18/2023 14:08:28 PDT):

>but both were custom licenses not MIT

Melanie Kambadur (8/18/2023 14:10:18 PDT):

Redacted

Nisha Deo (8/18/2023 14:20:01 PDT):

>My sense here is that the biggest news here is the assertion that Book3 was stripping books of copyright information -- rather than just the sheer number of copyrighted works they've apparently found

Nisha Deo (8/18/2023 14:20:12 PDT):

>because, otherwise the fact that Books3 allegedly has issues with copyrighted works is not new

Melanie Kambadur (8/18/2023 14:24:13 PDT):

>i feel like that's on the people who distributed the dataset somewhat, no?

Melanie Kambadur (8/18/2023 14:24:44 PDT):

>like by the time a piece of data gets put into a model for training it has been heavily processed, broken up into individual tokens (subwords)

Melanie Kambadur (8/18/2023 14:25:09 PDT):

>so obviously there is some point in model training that you need to not have copyright context but you don't necessarily distribute a datasource that way

Nisha Deo (8/18/2023 14:28:22 PDT):

>yeah that's a really good point. It's the piracy (and us knowing and being accomplices) that's the issue -- whether we knew or not at the time

Nisha Deo (8/18/2023 14:28:42 PDT):

>do we remove the copyright text or do we get the data like that anyways from there (silly Q)

Melanie Kambadur (8/18/2023 14:31:58 PDT):

>um to be honest sometimes we do move the copyright text as part of common cleaning because it is very repeated across datasets

Melanie Kambadur (8/18/2023 14:32:30 PDT):

>if you have a very repeated substring of data (such as a copyright notice), the model is prone to memorizing it and then generating it

Melanie Kambadur (8/18/2023 14:32:54 PDT):

>also in some circumstances it can be a waste of compute to train on duplicated data

Nisha Deo (8/18/2023 14:51:44 PDT):

>super helpful, I need to study up on all of this some more so I can catch these things more intuitively

Nisha Deo (8/18/2023 14:52:21 PDT):

>do you mind if I use some of this context in the comms plan explaining how things work?

Redacted

Redacted